Computational Statistics and Data Analysis 53 (2009) 3717–3725

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A generalized Waring regression model for count data

J. Rodríguez-Avi*, A. Conde-Sánchez, A.J. Sáez-Castillo, M.J. Olmo-Jiménez, A.M. Martínez-Rodríguez

Department of Statistics and Operations Research, University of Jaén, Spain

ARTICLE INFO

Article history:
Received 14 May 2008
Received in revised form 19 March 2009
Accepted 20 March 2009
Available online 28 March 2009

ABSTRACT

A regression model for count data based on the generalized Waring distribution is developed. This model allows the observed variability to be split into three components: randomness, internal differences between individuals and the presence of other external factors that have not been included as covariates in the model. An application in the field of sports illustrates its capacity for modelling data sets with great accuracy. Moreover, this yields more information than a model based on the negative binomial distribution.

1. Introduction

The Poisson distribution is used to model count data where the occurrence of an event is random and the occurrence rate is the same for all individuals. However, this occurrence rate may differ across individuals: this is known as heterogeneity (Long, 1997). When this heterogeneity may be explained by quantifiable and observable characteristics of the individuals, the Poisson regression model (*PRM*) is valid, in which the mean is a function of the observed variables or covariates. So, the heterogeneity is modelled as a deterministic function of the explanatory variables (Winkelmann, 2003). This model is the most basic for count data and is characterized by the equality of conditional mean and variance (equidispersion).

However, it is well known that the variability of data often exceeds the Poisson variability (overdispersion). It may be explained in several ways (Xekalaki, 1983, 2004; Winkelmann, 2003), among others, by the existence of unobserved heterogeneity, that is, by the presence of unfixed occurrence rates at each level of the model covariates. If it is assumed that this occurrence rate follows a gamma distribution, the resulting model is the negative binomial regression model (*NBRM*) (Hinde and Demétrio, 1998; Poortema, 1999; Cameron and Trivedi, 1998; Long, 1997). This allows us to consider a new source of variation that differs from the observed covariates and from randomness and thereby an additional component to explain the variability.

Moreover, the unobserved heterogeneity may be due to internal differences across individuals and external factors that might also be included as covariates in the regression model if they could be observed. In the *NBRM* both sources of variation in the occurrence rate are jointly considered by means of a gamma distribution.

In accident theory the univariate generalized Waring distribution, *UGWD*, (Irwin, 1968; Xekalaki, 1983) is an extension of the negative binomial distribution that allows three sources of variation to be distinguished: randomness, which is inherent in any random phenomenon, and the two aforementioned sources of heterogeneity between individuals, the one due to external factors, that is, different accident risk exposures (liability), and the other due to internal factors pertaining to each individual, that is, personal differences that are not related to external factors (proneness). A more general distribution which also considers this partition of the variance is studied by Rodríguez-Avi et al. (2007).

* Corresponding address: Despacho B3-058 Campus Universitario de Jaén, 23071 Jaén, Spain. Tel.: +34 953212207; fax: +34 953212034.
E-mail address: jravi@ujaen.es (J. Rodríguez-Avi).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.03.013

This work describes a regression model with a *UGWD* as its underlying distribution. The main advantage of this model over the *NBRM* is that the former allows us to distinguish the part of the unobserved heterogeneity due to the internal factors inherent to each individual and that due to the external factors such as those covariates that influence the variability of data but that have not been included in the model because they cannot be observed or measurable. From now on and by analogy with the terminology used in accident theory, these parts of the variance will be called proneness and liability, respectively. The performance of both models has been compared by simulation methods.

An example in the field of sports is considered to illustrate the behaviour of the model. Specifically, the dependent variable is the number of goals scored by the footballers of the Spanish football league over several seasons and the covariates are the position of the footballers on the pitch and the final classification of the team. Moreover, the number of matches played by each footballer has been initially considered only as offset and subsequently also as regressor. The effect of the covariates is studied, the fit obtained is compared against a regression model based on a negative binomial distribution and the relative weight of the three sources of variation (randomness, liability and proneness) in the presence of the covariates is computed.

2. Negative binomial regression models

Let Y be the response variable of a count model. In a *PRM*, $Y|x \sim Poisson(\lambda_x)$, where λ_x is the mean of the response variable for the values of the covariates, $x' = (x_1, \dots, x_p)$. Obviously, there is equidispersion in each level of the covariates, that is, $Var(Y|x) = E(Y|x)$.

As has been stated, if $Y|x$ is overdispersed, a way of explaining this excess variability is to propose a parametric model for λ_x. When $\lambda_x \sim Gamma(a_x, v_x)$, that is to say,

$$f(\lambda_x) = \frac{1}{v_x^{a_x}\Gamma(a_x)}\lambda_x^{a_x-1}\mathrm{e}^{-\lambda_x/v_x}, \quad \lambda_x > 0, a_x, v_x > 0.$$

$Y|x$ has a negative binomial distribution with probability mass function (p.m.f.)

$$f(y|x) = \frac{\Gamma(a_x+y)}{\Gamma(a_x)y!}\left(\frac{1}{1+v_x}\right)^{a_x}\left(\frac{v_x}{1+v_x}\right)^{y}, \quad y = 0, 1, 2, \dots, a_x, v_x > 0, \tag{1}$$

denoted by $Y|x \sim NB(a_x, p_x)$ with $p_x = (1+v_x)^{-1}$. In this case, an *NBRM* arises. It should be emphasized that the model about λ_x is related to the unexplained heterogeneity, independently of its origin. In this model it verifies that

$$E(Y|x) = E(E(Y|x, \lambda_x)) = E(\lambda_x) = \mu_x = a_x v_x.$$

Different *NBRM* can be generated by linking μ_x, a_x and v_x with the explanatory variables. One of the most usual in data processing is given by

$$a_x = a, \qquad \mu_x = \mathrm{e}^{\beta_0 + x'\beta},$$

with $\beta' = (\beta_1, \dots, \beta_p)$, where a does not depend on the covariates but does v_x. This model is known as *Negbin* II (Cameron and Trivedi, 1986) and it establishes a linear variance-mean rate:

$$\begin{aligned} Var(Y|x) &= E(Var(Y|x, \lambda_x)) + Var(E(Y|x, \lambda_x)) \\ &= E(\lambda_x) + Var(\lambda_x) = \mu_x + \frac{1}{a}\mu_x^2 = \mu_x\left(1 + \frac{1}{a}\mu_x\right). \end{aligned}$$

The first term represents the variability due to randomness and the second to differences between individuals. The partition of the variance for this model appears in Table 1. It can be observed that the variance rate due to heterogeneity across individuals tends to 1 as μ_x increases, whereas the variance rate due to randomness tends to 0. This means that, as the average number of occurrences of an event increases, the observed variability is more due to the individual heterogeneity than to randomness.

It should be pointed out that if $a \to \infty$ and $v_x \to 0$ with μ_x constant, the *Negbin* II model tends to the *PRM*, so the latter is nested within the former.

On the other hand, if v does not depend on the covariates but does a_x, that is, $v_x = v$ and $\mu_x = \mathrm{e}^{\beta_0 + x'\beta}$, the *Negbin* I model appears (Cameron and Trivedi, 1986) in which the variance-mean rate is constant:

$$Var(Y|x) = (1+v)\mu_x.$$

In order to make comparisons, we focus on the *Negbin* II model, since it provides better fits for data included here than does the *Negbin* I model.

The NBRM recently appears within more general frameworks to model count data variables, as in Rigby et al. (2008), where all the distribution parameters can be modelled as functions of explanatory variables, or in Cordeiro et al. (2009), where a new class of discrete generalized nonlinear models is introduced.

Table 1
Partition of the variance in the *Negbin* II model.

Source of variability	Variance	Variance rate
Randomness	μ_x	$\frac{a}{a+\mu_x}$
Heterogeneity across individuals	$\frac{1}{a}\mu_x^2$	$\frac{\mu_x}{a+\mu_x}$
Total	$\mu_x + \frac{1}{a}\mu_x^2$	1

3. Generalized Waring regression model

In accident theory, Irwin (1968) criticizes the fact that the *NB* distribution considers jointly the effects of liability and proneness. Hence, he proposes the *UGWD* in such a way that the gamma distribution models one of these sources of variation (liability) and introduces a beta distribution for the other one (proneness).

In the case of the regression model under discussion, we have extended Irwin's methodology by introducing covariates as explanatory variables into the model. Specifically, it is assumed that proneness does not depend on the covariates because it contains all the inherent conditions to each individual, so the distribution of the proneness is the same for all levels of the covariates. Hence, it has sense to consider liability conditional on proneness but not on the contrary, since liability represents the heterogeneity due to those covariates that have not been included in the model because they are not observable or measurable. Then, if $x' = (x_1, \ldots, x_p)$ is the vector of covariates, let v be the proneness and $\lambda_x|v$ the liability for a given proneness. So, the following hypotheses are considered:

1. $Y|x, \lambda_x, v \sim Poisson(\lambda_x)$.
2. $\lambda_x|v \sim Gamma(a_x, v)$. Therefore, $Y|x, v \sim NB(a_x, p)$ with $p = \frac{1}{1+v}$.
3. $v \sim BetaII(\rho, k)$, that is,

$$f(v) = \frac{\Gamma(k+\rho)}{\Gamma(k)\Gamma(\rho)} v^{k-1}(1+v)^{-(k+\rho)}, \quad v > 0,\ k, \rho > 0,$$

or, equivalently, the probability *p* of increasing the variable changes from individual to individual by means of a $Beta(\rho, k)$ distribution.

In this case, the p.m.f. of $Y|x$ is (Irwin, 1968; Xekalaki, 1983)

$$f(y|x) = \frac{\Gamma(a_x+\rho)\Gamma(k+\rho)}{\Gamma(\rho)\Gamma(a_x+k+\rho)} \frac{(a_x)_y (k)_y}{(a_x+k+\rho)_y} \frac{1}{y!}, \quad y = 0, 1, 2, \ldots, \tag{2}$$

corresponding to a $UGWD(a_x, k, \rho)$, where $a_x, k, \rho > 0$ and $(\alpha)_r = \frac{\Gamma(\alpha+r)}{\Gamma(\alpha)}$ if $\alpha > 0$.

4. By analogy with the classical regression models and for simplicity, we impose the equation of log-linearity for the mean

$$E(Y|x) = \mu_x = e^{\beta_0 + x'\beta} = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}$$

with $\beta' = (\beta_1, \ldots, \beta_p)$.

Furthermore, the mean of the *UGWD* is given by

$$E(Y|x) = \frac{a_x k}{\rho - 1},$$

so, ρ must be greater than 1 (if $0 < \rho < 1$ the distribution has an infinite mean), and then

$$a_x = \frac{\mu_x(\rho-1)}{k}.$$

5. Finally, in order to guarantee the conditions $k > 0$ and $\rho > 1$, these parameters are written as

$$k = e^{k_0}, \qquad \rho - 1 = e^{\rho_0},$$

with $k_0, \rho_0 \in \mathbb{R}$.

From now on, the model will be called the generalized Waring regression model (*GWRM*). The partition of the variance in this regression model is established as:

$$\begin{aligned} Var(Y|x) &= E(Var(Y|x, v)) + Var(E(Y|x, v)) \\ &= E(a_x v + a_x v^2) + Var(a_x v) \\ &= E(a_x v) + E(a_x v^2) + Var(a_x v) \\ &= \frac{a_x k}{\rho-1} + \frac{a_x k(k+1)}{(\rho-1)(\rho-2)} + \frac{a_x^2 k(k+\rho-1)}{(\rho-1)^2(\rho-2)} \\ &= \mu_x + \frac{k+1}{\rho-2}\mu_x + \frac{k+\rho-1}{\rho-2}\frac{\mu_x^2}{k}, \end{aligned} \tag{3}$$

for $a_x, k > 0$ and $\rho > 2$ (if $\rho < 2$, the distribution has infinite variance).

Table 2
Partition of the variance in the *GWRM*.

Source of variability	Variance	Variance rate
Randomness	μ_x	$\frac{\rho-2}{k+\rho-1}\,\frac{k}{k+\mu_x}$
Liability	$\frac{k+1}{\rho-2}\mu_x$	$\frac{k+1}{k+\rho-1}\,\frac{k}{k+\mu_x}$
Proneness	$\frac{\rho+k-1}{\rho-2}\,\frac{\mu_x^2}{k}$	$\frac{\mu_x}{k+\mu_x}$
Total	$\frac{k+\rho-1}{\rho-2}\left(\mu_x+\frac{1}{k}\mu_x^2\right)$	1

The first term of this decomposition represents the variability due to randomness and comes from the underlying Poisson model. The other two terms refer to the variability that is not due to randomness but is explained by the presence of liability and by proneness, respectively. This decomposition appears in Table 2. In this case the variance rates due to randomness and liability decrease as μ_x increases, whereas the variance rate due to proneness increases.

Moreover, it can be observed that the variance function of a *GWRM* is a combination of the *Negbin* I and *Negbin* II variance functions given by

$$Var(Y|x) = \phi\left(\mu_x + \frac{1}{k}\mu_x^2\right), \tag{4}$$

with $\phi = (k+\rho-1)/(\rho-2)$.

There is an important question about the identification of the non-random variance components (liability and proneness). One of the main drawbacks of using the $UGWD(a, k, \rho)$ is that the parameters a and k are interchangeable when there is no auxiliary information given by the covariates. This fact may be confirmed in (2). This identification problem prevents liability and proneness components from being distinguished in the univariate fits. To solve it, Irwin (1968) proposed that the expert should deduce which of these components is which from their own knowledge of the phenomenon. Xekalaki (1984) proposed a less subjective solution, developing a bivariate model that divides the observation period into two non-overlapping subperiods in which the model for proneness does not change. In the proposed *GWRM* with, at least, one covariate, the parameters a and k are not interchangeable because, as in Xekalaki's bivariate model, the random model for proneness does not change. So, the identification problem of the non-random components is solved.

The parameter estimation in this model has been carried out by the direct optimization of the log-likelihood obtained from Eq. (2). Specifically, the *nlm* and *optim* functions of R (Team, 2007) have been used to maximize the log-likelihood function obtained from (2). These functions are based on Nelder–Mead, quasi-Newton and conjugate-gradient algorithms. In practice, several initial values have been used to guarantee the convergence to the global optimum. Differences in the estimates have been detected only when these estimates are in the boundary of the parametric space, because they are usually very high.

Irwin (1968) shows that the UGWD converges to the NB distribution. Similarly the GWRM converges to the NBRM in two ways:

- Firstly, if $k, \rho \to \infty$ with the same order of convergence

$$\begin{aligned} f(y|x) &\propto \frac{(a_x)_y}{y!}\,\frac{(\theta(\rho-1))_y}{(a_x+(1+\theta)\rho-\theta)} \\ &= \frac{(a_x)_y}{y!}\,\frac{\theta^y\rho^y+O(\rho^{(r-1)})}{(1+\theta)^y\rho^y+O(\rho^{(r-1)})} \to \frac{(a_x)_y}{y!}\left(\frac{\theta}{1+\theta}\right)^y, \end{aligned}$$

  that is the kernel of the *NB* $\left(a_x, \frac{1}{1+\theta}\right)$ density, where $\theta = \frac{k}{\rho-1}$. Let us observe that $\mu_x = a_x\theta$ and, from (4), $Var(Y|x) = \mu_x(1+\theta)$. Hence, the variance is a linear function of the mean, and the *Negbin* I model is obtained.
- Similarly, if $\rho \to \infty$ and $\theta_x = \mu_x/k$ is bounded, then $a_x \to \infty$ with the same order of convergence and

$$\begin{aligned} f(y|x) &\propto \frac{(k)_y}{y!}\,\frac{(\theta_x(\rho-1))_y}{(k+(1+\theta_x)\rho-\theta_x)} \\ &= \frac{(k)_y}{y!}\,\frac{\theta_x^y\rho^y+O(\rho^{(r-1)})}{(1+\theta_x)^y\rho^y+O(\rho^{(r-1)})} \to \frac{(k)_y}{y!}\left(\frac{\theta_x}{1+\theta_x}\right)^y, \end{aligned}$$

  that is the kernel of the *NB* $\left(k, \frac{1}{1+\theta_x}\right)$ density. Thus, from (4), $Var(Y|x) = \mu_x\left(1+\frac{\mu_x}{k}\right)$, corresponding to a *Negbin* II model.

From these convergence results it can be inferred that the *Negbin* models are nested in the *GWRM*. Nevertheless, the *Negbin* models appear at the edge of the parameter space of the *GWRM*, which is a problem with the likelihood ratio test (LRT) for comparing both models, since under the null hypothesis the true parameter lies on the boundary of the parameter space.

Table 3
Means and standard errors (below in brackets) of estimates in *GWRM* and *NBRM* fits of *GWRM* generated data with $k = 2.5$, $\rho = 3.5$, $\beta_0 = 1.25$ and $\beta_1 = 1$.

		$N = 100$	$N = 300$	$N = 500$
GWRM	$\hat{k}$	6.48 (10.89)	4.61 (5.75)	3.94 (3.56)
	$\hat{\rho}$	4.50 (2.77)	4.26 (1.56)	4.07 (1.07)
	$\hat{\beta}_0$	1.24 (0.22)	1.25 (0.13)	1.25 (0.11)
	$\hat{\beta}_1$	1.01 (0.38)	0.98 (0.21)	0.99 (0.17)
NBRM	$\hat{a}$	1.11 (0.25)	1.04 (0.13)	1.04 (0.10)
	$\hat{\beta}_0$	1.22 (0.26)	1.25 (0.16)	1.24 (0.13)
	$\hat{\beta}_1$	1.02 (0.46)	1.00 (0.27)	1.00 (0.21)
LRT	*p-values* < 0.01	709	969	998

In this case the asymptotic distribution of the LRT statistic has probability mass of one half at zero and a half-$\chi^2(1)$ distribution above 0. So, if testing at level $\alpha > 0.5$, one rejects H_0 if the test statistic exceeds $\chi^2_{1-2\alpha}$ rather than $\chi^2_{1-\alpha}$ (Cameron and Trivedi, 1998).

4. Simulation

A study based on simulation has been carried out in order to illustrate some aspects related to the behaviour of the *GWRM* model and the estimation process of its parameters.

Firstly, $s = 1000$ samples of sizes $N = 100, 300, 500$ have been simulated under a *GWRM* model for a specific choice of its parameters. Secondly, the *GRWM* and *Negbin* II models have been fitted for each sample and the results obtained have been compared. Similarly, $s = 1000$ samples of sizes $N = 100, 300, 500$ have been simulated under a *Negbin* II model. These samples have been also modelled by the *GRWM* and the *Negbin* II models.

4.1. Simulation under a GWRM model

Data are generated from a *GWRM* model with one regressor. Specifically, $\mu_i = \exp(\beta_0 + \beta_1 x_i)$ where x_i $(i = 1, \ldots, N)$ is generated from the uniform distribution on the unit interval. The selected values of the parameters have been $k = 2.5$, $\rho = 3.5$, $\beta_0 = 1.25$ and $\beta_1 = 1$, so that the values of liability and proneness are not too low. Thus, the *i*th datum of each sample is simulated by a $UGWD(a_i, k, \rho)$, where $a_i = \mu_i(\rho - 1)/k$, using the *rghyper* function of the *SuppDist* R package: this function simulates values of a $UGWD(a, k, \rho)$ as a type IV generalized hypergeometric distribution with parameters $-k$, a and $\rho - 1$ (Kemp and Kemp, 1956). All the simulations use the same draw $x_1, \ldots, x_N$ and are performed $s = 1000$ times. The parameters to be estimated are $k_0 = \log(k)$, $\rho_0 = \log(\rho - 1)$, β_0 and β_1.

The LRT shows that some of the *GWRM* fits are not significantly better than those obtained for the *Negbin* II model (see last row in Table 3). In these cases, the parameter estimates seem to indicate the convergence of the *GWRM* to the *Negbin* II model. This happens in 291 of 1000 fits for $N = 100$, in 31 when $N = 300$ and in 2 when $N = 500$. As a first conclusion it can be said that the sample size influences the identification of the model. Moreover, when $N = 100$, there are 9 *GWRM* fits, significantly better than for the *Negbin* II model, that have high estimates of ρ (> 100); this also suggests the convergence to a *Negbin* II model.

Table 3 also includes the estimates of the parameters and their standard errors after removing the fits aforementioned. In spite of this previous filter, one of the main drawbacks of the *GWRM* model can be appreciated: the estimates of k and ρ are very biased and dispersed, which is more evident as the sample size decreases. Fig. 1 shows the histograms of the estimates of k and ρ for $N = 100$, 300 and 500. These indicate that the distributions of the parameter estimators of k and ρ have an increasing right asymmetry as the sample size decreases.

On the other hand, the estimates of β_0 and β_1 have unimportant bias, low standard errors and normal shapes.

4.2. Simulation under a Negbin II model

In this case, $s = 1000$ samples with $N = 100, 300, 500$ data have been generated from a *Negbin* II model with the same regressor x, $a = 1.5$ and the same parameters $\beta_0 = 1.25$ and $\beta_1 = 1$.

The means, the standard errors of the parameter estimates and the number of fits in which the *Negbin* II model is significantly better than the corresponding *GWRM*, according to the LRT, appear in Table 4. It can be emphasized that the results of the LRT as well as the estimated values of ρ seem to indicate the convergence of the *GWRM* model to the *Negbin* II model. However, problems in the estimation of k arise when $N = 100$: 76 estimates are greater than 1000, so the average estimate of k decreases until 25 excluding them. If $N = 300$, only 145 estimates of k are greater than 10 and the new average

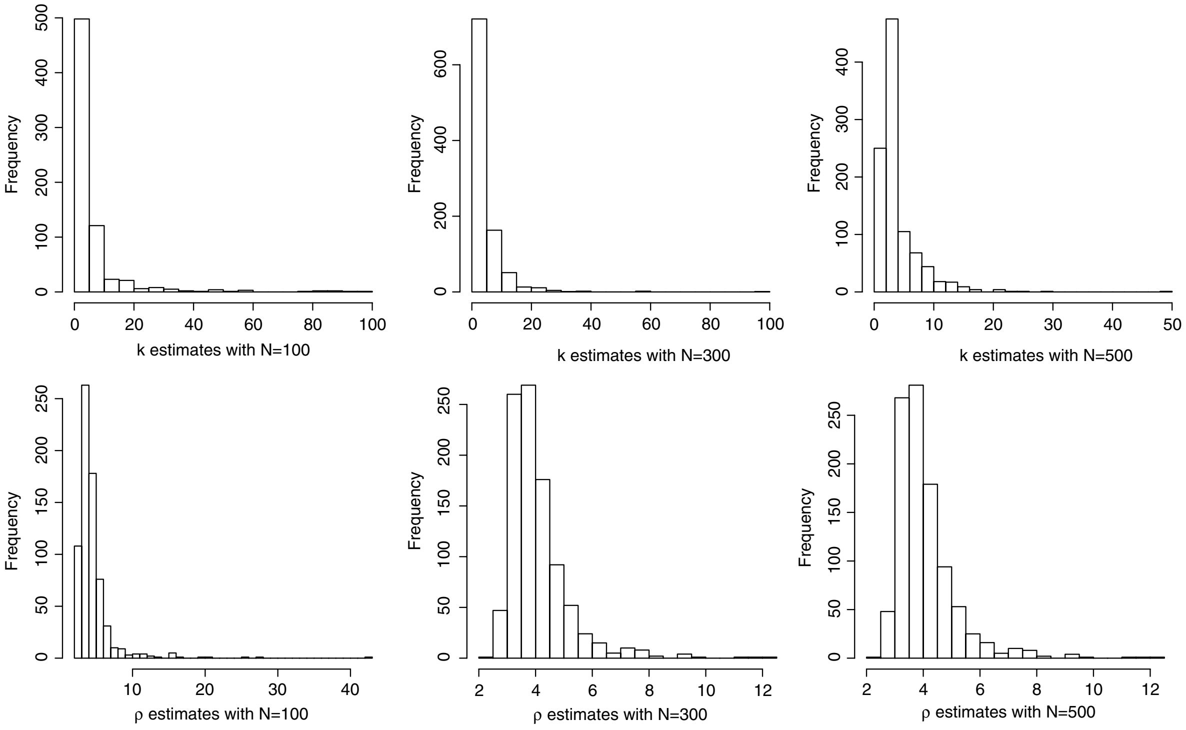

Fig. 1. Histograms of the estimates of k and ρ for each sample size and data generated by a *GWRM* model.

Table 4
Means and standard errors (below in brackets) of estimates in *GWRM* and *NBRM* fits of *NBRM* generated data with $a = 1.5$, $\beta_0 = 1.25$ and $\beta_1 = 1$.

		$N = 100$	$N = 300$	$N = 500$
GWRM	$\hat{k}$	2.39×10^5 (5.07×10^6)	7.81 (17.43)	1.74 (1.97)
	$\hat{\rho}$	1.05×10^{11} (1.55×10^{12})	6.42×10^{13} (2.03×10^{15})	7.03×10^{12} (9.76×10^{12})
	$\hat{\beta}_0$	1.26 (0.22)	1.27 (0.11)	1.25 (0.09)
	$\hat{\beta}_1$	0.98 (0.33)	0.97 (0.19)	0.99 (0.15)
NBRM	$\hat{a}$	1.58 (0.31)	1.53 (0.17)	1.52 (0.12)
	$\hat{\beta}_0$	1.24 (0.18)	1.26 (0.11)	1.25 (0.08)
	$\hat{\beta}_1$	1.01 (0.33)	0.99 (0.19)	0.99 (0.15)
LRT	*p-values* < 0.01	163	104	101

estimate, 1.67, is much nearer the true value of $a = 1.5$. This situation would be the most desirable since the *GWRM* would converge to the *Negbin* II model used to simulate the samples. If $N = 500$, this convergence is clearer: the estimates of k are next to the value of $a = 1.5$ whereas the estimates of ρ are, in general, very high.

5. Practical application

From a statistical point of view, the number of goals scored by a footballer in a season is a discrete count variable with a clear tendency towards overdispersion. This excess variability by comparison with the Poisson model, which only considers the effect of pure chance, may be due to several causes:

- A set of external factors observable by covariates that significantly influence the risk of scoring goals; among them, for instance, the footballers position in the field or the quality of the team, that may be measured by its final classification. These and other covariates would determine what in accident theory is called liability.
- When it comes to scoring a goal there is a factor that is not related to environmental/external factors but is associated with the characteristics of each player, that is, with their goal-scoring ability and intelligence. Indeed, in the football world, this component is known as a footballer's *goal-scoring intuition*, and may be comparable to proneness in accident theory.

Table 5
Goodness of fits.

Covariates	Model	Log-likelihood	AIC
x_1, x_2, x_3	*Negbin* II	−1897.069	3804.1
	GWRM	−1883.983	3775.967
$x_1, x_2, x_3, x_1x_2, x_1x_3$	*Negbin* II	−1895.016	3804.0
	GWRM	−1880.358	3772.717
x_1, x_2, x_3, log (offset)	*Negbin* II	−1823.564	3659.1
	GWRM	−1819.833	3649.667
x_1, x_2, x_3, log (offset), x_1x_2, x_1x_3	*Negbin* II	−1822.369	3660.7
	GWRM	−1818.139	3650.279

Table 6
Maximum likelihood estimates and standard errors for the *GWRM* without interaction.

Covariates	Estimates	s.e.	z
(Intercept)	−5.703	0.296	−19.241
x_1	−0.027	0.005	−5.350
x_2	0.771	0.083	9.278
x_3	1.725	0.084	20.322
log(offset)	0.855	0.082	10.337

In this context, we think that the *GWRM* proposed may be capable of quantifying the importance of these three sources of variability, providing more information about data than any other regression model based on the negative binomial distribution.

The response variable is Y, the number of goals scored by the footballers in the first division of the Spanish league. Data have been obtained from the sports paper *MARCA* and are referred to the 2000/2001 to 2006/2007 seasons. Since there are footballers who play more than one season, we have selected the season in which each one has played more matches. Thus, we try to guarantee the independence of data. The population is composed of 1224 footballers, excluding goalkeepers.

The covariates considered are:

- The final classification of the team in each season, x_1, with values from 1 to 20.
- The position in the field, with three levels (forward, midfielder and defender), coded by two dummy variables, x_2 and x_3, with defender as the reference category. This is the main position considered by the sports paper *MARCA*.

Furthermore, the variable number of matches played, with values from 1 to 38, has been included as an offset. This means that a doubling in exposure time (the number of matches played) doubles the expected count (the number of goals scored). If this proportionality assumption is given free for test, the offset may be included as a regressor without restricting its coefficient to unity. Here, the equation considered for the mean is

$$\mu_x = \exp(\beta_0 + x'\beta + \gamma \log(\mathit{offset}) + \log(\mathit{offset})),$$

so that the logarithm of the number of matches played is included twice, both as offset and as regressor. The test for proportionality simplifies to testing $H_0 : \gamma = 0$ (Winkelmann, 2003).

So, the model assumes that proneness represents the overdispersion due to between-player variation in goal-scoring ability for players in the same position in the same team, while liability represents the overdispersion due to missing covariates which would affect players in the same position in the same team identically.

The *Negbin* II model and the *GWRM* have been fitted including both covariates without interactions (x_1, x_2, x_3) and with interactions ($x_1, x_2, x_3, x_1x_2, x_1x_3$). In both the cases the logarithm of the offset has also been considered as regressor.

To evaluate the accuracy of the fits we have computed the value of the log-likelihood function and the AIC (see Table 5). It can be observed that the best fit is provided by the *GWRM* in both the cases. The LRT has been used to test whether the interaction between the covariates contributes significantly to the response variable, concluding that this interaction is not relevant. This test has also allowed us to prove that the regressor log(offset) is significant in the model.

Table 6 contains the maximum likelihood parameter estimates for the best fitted model (*GWRM* without interaction and log(offset) as regressor) and their respective standard errors. The estimates of k_0 and ρ_0 are 1.5104 and 2.9008, respectively, and their standard errors 0.3207 and 0.2921. Therefore, $\hat{k} = 4.5284$ and $\hat{\rho} = 19.1882$. Comparing this fit with that obtained by the *Negbin* II model, the value of the likelihood ratio statistic is 7.4611 and the associated $p - value$, with one degree of freedom, 0.003.

In Table 7 expected frequencies are compared with observed frequencies. It can be noticed that the expected frequencies for the *GWRM* are nearer to the observed frequencies than those computed for the *NBRM*.

Starting from these estimates the increments in the average number of goals scored (odds ratios) have been obtained according to the classification of the football team and the position of the footballer in the pitch, taking the *defenders* as the reference category (Table 8). Taking into account that $e^{\beta_j} - 1$ is the relative effect of a unit change in x_j on the expected value of y, it can be observed that the average number of goals scored by a footballer decreases by 2.7347% as the team loses one

Table 7
Observed (O_i) and expected (E_i) global frequencies for the (a) *NBRM* and (b) *GWRM*.

y_i	O_i	$E_i(a)$	$E_i(b)$	y_i	O_i	$E_i(a)$	$E_i(b)$
0	550	517.910	544.688	15	6	3.634	3.624
1	205	219.542	206.799	16	2	2.972	2.978
2	143	139.358	132.994	17	4	2.437	2.458
3	80	93.163	89.497	18	1	2.002	2.035
4	56	63.970	61.885	19	3	1.646	1.690
5	39	45.046	43.846	20	2	1.354	1.408
6	32	32.510	31.791	21	0	1.115	1.175
7	18	24.008	23.555	22	2	0.917	0.982
8	21	18.099	17.799	23	1	0.755	0.823
9	20	13.895	13.688	24	4	0.621	0.690
10	9	10.834	10.687	25	0	0.511	0.580
11	12	8.558	8.453	26	0	0.420	0.488
12	5	6.833	6.759	27	0	0.345	0.411
13	4	5.503	5.453	28	0	0.283	0.347
14	4	4.461	4.432	≥ 29	1	1.298	1.985

Table 8
Expected odds ratios.

Covariates	e^{β_j}
Classification	0.972
Midfielder	2.162
Forward	5.617

position in the final classification. Also, the average number of goals scored by a midfielder is 2.162 times greater than the defender, whereas this average is 5.617 times greater for a forward. Furthermore, the coefficient of the offset indicates that the average number of goals increases by 1.855% when the number of matches played increases by 1%.

Finally, Fig. 2 shows the fractions of variance attributed to randomness, liability and proneness for all the values of the covariates. It may be seen that:

- Randomness and liability decrease as the number of matches played increases and improves the classification of the team, whereas proneness increases.
- Proneness is more important the nearer the footballer is to the goal, since the forwards have greater proneness than the midfielders who, in turn, have greater proneness than the defenders.
- Similarly, defenders have the greatest percentage of randomness, followed by midfielders and forwards.

6. Discussion

A count data regression model offers a richer set of interesting inferences, such as the computing of marginal probability effects in order to trace the response of the entire count distribution to small changes in explanatory variables. Specifically, the proposed model offers additional interesting interpretations of the data generating process that were not possible with simpler models.

In this sense, the *GWRM* proposes a *UGWD* for each level combination of the observed covariates that allows the variability of the dependent variable to be explained by three different sources:

- Randomness, attached to any stochastic phenomenon.
- Variability given by external factors or liability, due to the presence of other covariates that cannot be specified and that influence the model. In the practical application, for instance, there can be differences due to the position of the forwards that cannot be quantified by the variable *position*.
- Variability given by internal factors inherent to each individual or proneness. Beyond measurable conditions (covariates) and other factors (liability), there are characteristic conditions of each individual that affect the response variable. In the practical application, we are sure that besides the characteristics of the team, the position in the pitch, the number of matches played, etc, there are characteristics no measurable of each footballer that influence the number of goals scored.

Another advantage of the *GWRM* is that it allows the non-random components of the variance to be identified because the parameters a and k are not interchangeable any more, without needing to suggest a bivariate model.

Moreover, the example and the simulation study show that the *GWRM* allows us to obtain equal or better fits than the *NBRM*, because the latter may be seen as a limit case of the former in the boundary of the parametric space. However, the main drawback is the identification of the model because the parameters k and ρ are highly correlated, so the estimates obtained are sometimes biased.

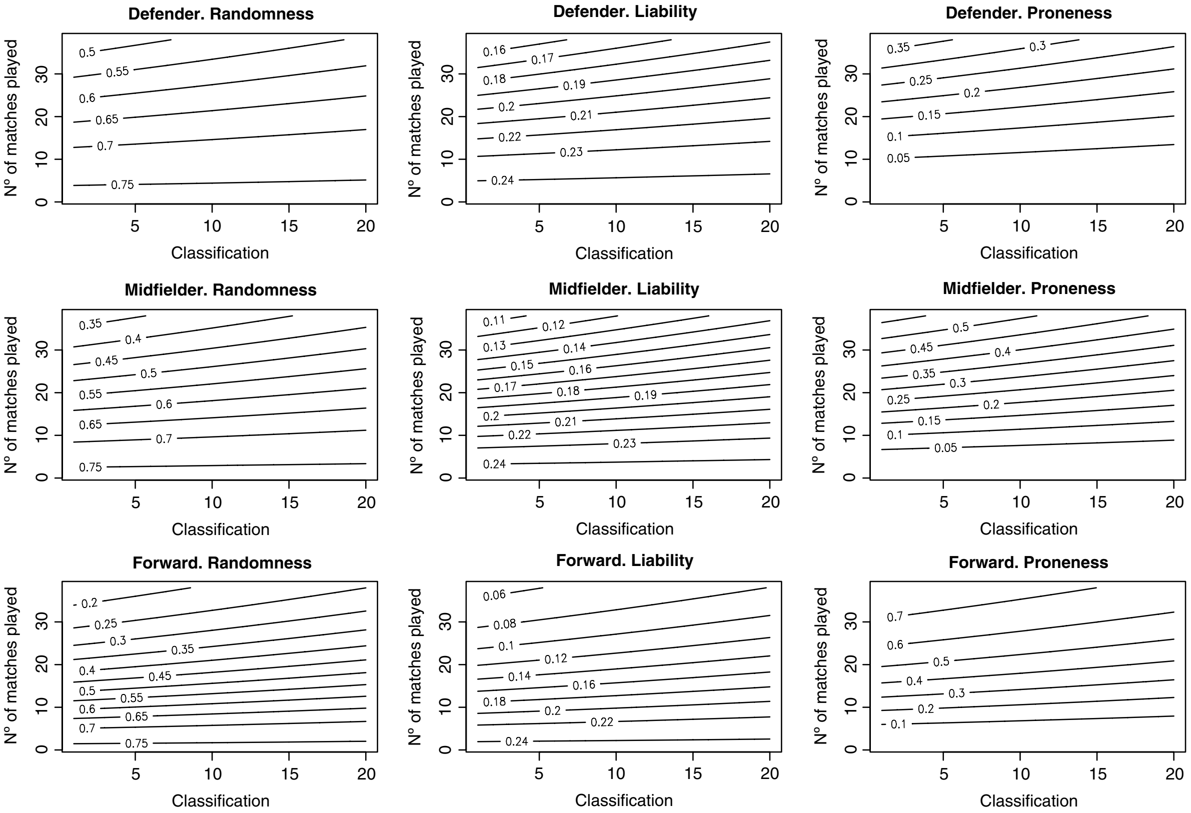

Fig. 2. Contour plots of variance component fractions.

The practical application has revealed that the model can explain specific characteristics of the studied phenomenon. Thus, we have confirmed that the conclusions obtained about the variability of data coincide with our previous conception about this phenomenon.

Therefore, the utility of the model is justified in those cases in which liability and proneness may appear separately. But if this does not occur, the simulation study has shown that the same model is valid for detecting the presence of an only source of unobserved heterogeneity, because in the simulations under a *Negbin* II model, most of the *GWRM* fits converged to this model.

Acknowledgements

The authors wish to thank the referees for their many helpful comments and suggestions, which have substantially improved the manuscript.

References

Cameron, A., Trivedi, P.K., 1986. Econometric models based on count data: comparisons and applications of some estimators and tests. Journal of Applied Econometrics 1, 29–53.
Cameron, A., Trivedi, P.K., 1998. Regression Analysis of Count Data. In: Econometric Society Monographs, vol. 30. Cambridge University Press, Cambridge.
Cordeiro, G.M., Andrade, M.G., de Castro, M., 2009. Power series generalized nonlinear models. Computational Statistics and Data Analysis 53, 1155–1166.
Hinde, J., Demétrio, C.G.B., 1998. Overdispersion: Models and estimation. Computational Statistics and Data Analysis 27, 151–170.
Irwin, J.O., 1968. The generalized waring distribution applied to accident theory. Journal of the Royal Statistical Society. Series A 131 (2), 205–225.
Kemp, C.D., Kemp, A.W., 1956. Generalized hypergeometric distributions. Journal of the Royal Statistical Society, Series B 18, 202–211.
Long, J.S., 1997. Regression Models for Categorical and Limited Dependent Variables. SAGE Publications, Thousand Oaks.
Poortema, K., 1999. On modelling overdispersion of counts. Statistica Neerlandica 53 (1), 5–20.
Rigby, R., Stasinopoulos, D., Akantziliotou, C., 2008. A framework for modelling overdispersed count data, including the poisson-shifted generalized inverse gaussian distribution. Computational Statistics and Data Analysis 53, 381–393.
Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A.J., Olmo-Jiménez, M.J., 2007. A new generalization of the waring distribution. Computational Statistics and Data Analysis 51 (12), 6138–6150.
R Development Core Team, 2007. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.
Winkelmann, R., 2003. Econometric Analysis of Count Data. Springer, Berlin.
Xekalaki, E., 1983. The univariate generalized waring distribution in relation to accident theory: Proneness, spells or contagion. Biometrics 39 (4), 887–895.
Xekalaki, E., 1984. The bivariate generalized waring distribution and its application to accident theory. Journal of the Royal Statistical Society. Series A 147 (3), 488–498.
Xekalaki, E., 2004. Under and over dispersion. Encyclopedia of Actuarial Science 3, 1700–1705.